Московский государственный университет
имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики

Отчет по заданию №2

Выполнили:
Данько Артем
Ковалева Василина
Виль Вадим

Преподаватель:
Гусева Юлия

Москва
2018

Содержание

Постановка задачи

1. Считать данные из training.xlsx. Ответы на тестовой выборке testing.xlsx не следует использовать ни в каких экспериментах, кроме финального. Проверить является ли ряд стационарным в широком смысле. Это можно сделать двумя способами

   - Провести визуальную оценку, отрисовав ряд и скользящую статистику(среднее, стандартное отклонение). Постройте график на котором будет отображен сам ряд и различные скользящие статистики.
   - Провести тест Дики - Фуллера.
   - Сделать выводы из полученных результатов. Оценить достоверность статистики. (25 баллов)

2. Разложить временной ряд на тренд, сезональность, остаток в соответствии с аддитивной, мультипликативной моделями. Визуализировать их, оценить стационарность получившихся рядов, сделать выводы. (15 баллов)

3. Проверить является ли временной ряд интегрированным порядка k. Если является, применить к нему модель $ARIMA$, подобрав необходимые параметры с помощью функции автокорреляции и функции частичной автокорреляции. Выбор параметров обосновать. Отобрать несколько моделей. Предсказать значения для тестовой выборки. Визуализировать их, посчитать r2 score для каждой из моделей. Произвести отбор наилучшей модели с помощью информационного критерия Акаике. Провести анализ получившихся результатов. (50 баллов)

- За все правильно выполненные пункты можно получить 90 баллов.
- +10 баллов - соблюдение PEP8
- +10 баллов - использование для визуализации библиотек bokeh или seaborn.
- Надо сделать, чтобы было красиво.

Решение

Вводные понятия

Под временным рядом понимаются последовательно измеренные через некоторые (зачастую равные) промежутки времени данные. Прогнозирование временных рядов заключается в построении модели для предсказания будущих событий, основываясь на известных событиях прошлого.

Анализ временных рядов — совокупность математико-статистических методов анализа, предназначенных для выявления структуры временных рядов и для их прогнозирования.

Дисперсия выборки - это среднее арифметическое квадратов отклонений. Отклонение - это разность числа и некоторой точки отчёта, чаще всего это среднее арифметическое или медиана. Например, если у нас есть следующий ряд чисел: 1; 2; 3; 4; 5; 6; 7, то его среднее арифметическое это сумма чисел ряда, деленная на их количество, то есть $(1 + 2 + 3 + 4 + 5 + 6 + 7) : 4 = 28 : 7 = 4$ (здесь среднее арифметическое набора чисел совпадает с медианой). Тогда найдём отклонения. Они будут соответственно -3; -2; -1; 0; 1; 2; 3. Тогда квадраты отклонений будут 9; 4; 1; 0; 1; 4; 9. Найдём их среднее арифметическое: $(9 + 4 + 1 + 0 + 1 + 4 + 9) : 7 = 28 : 7 = 4$. Получаем, что дисперсия данного набора равна 4. Ковариация (корреляционный момент, ковариационный момент) — в теории вероятностей и математической статистике мера линейной зависимости двух случайных величин.

Ряд называется слабо стационарным или стационарным в широком смысле, если его среднее значение и дисперсия не зависят от времени, а ковариационная функция зависит только от сдвига. Если нарушается хотя бы одно из этих условий, то ряд является нестационарным.

Тест Дики — Фуллера — это методика, которая используется в прикладной статистике и эконометрике для анализа временных рядов для проверки на стационарность. Является одним из тестов на единичные корни (Unit root test).

Временной ряд имеет единичный корень, или порядок интеграции один, если его первые разности образуют стационарный ряд.

Как и большинство других видов анализа, анализ временных рядов предполагает, что данные содержат систематическую составляющую (обычно включающую несколько компонент) и случайный шум (ошибку), который затрудняет обнаружение регулярных компонент. Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей.

Таким образом, каждый уровень временного ряда может формироваться из трендовой (T), циклической или сезонной компоненты (S), а также случайной (E) компоненты.

Модели, где временной ряд представлен в виде суммы перечисленных компонентов называются аддитивными, если в виде произведения – мультипликативными моделями.

Аддитивная модель имеет вид: Y = T + S + E

Мультипликативная модель имеет вид: Y = T * S * E

Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени. Сезонность – строго периодические и связанные с календарным периодом отклонения от тренда:

- Аддитивная сезонность – амплитуда сезонных колебаний не имеет ярко выраженной тенденции к изменению во времени.
- Мультипликативная сезонность – амплитуда сезонных колебаний имеет выраженную тенденцию к изменению во времени.

Скользящая статистика — общее название для семейства функций, значения которых в каждой точке определения равны среднему значению исходной функции за предыдущий период. Скользящая статистика обычно используются с данными временных рядов для сглаживания краткосрочных колебаний и выделения основных тенденций или циклов.

Суть метода

1. Скользящая статистика:
   - Среднее

     В общем случае, взвешенные скользящие средние вычисляются по формуле:

     $$WWMA_t = \frac{1}{n} * \sum_{i=0}^{n-1} p_{t-i} \tag{1}$$

     , где n - окно скользящей статистики.
   - Стандартное отклонение

     $$SD_t = \sqrt{\frac{1}{n} * \sum_{i=0}^{n-1} (p_{t-i} - ME)^2} \tag{2}$$

     , где n - окно скользящей статистики, а ME - среднее.
2. Тест Дики — Фуллера: Рассмотрим авторегрессионнное уравнение:

   $$y_t = a * y_{t-1} + \varepsilon_t \tag{3}$$

   , где y_t — временной ряд, а ε — ошибка.

   Если $a = 1$, то процесс имеет единичный корень, в этом случае ряд y_t не стационарен, является интегрированным временным рядом первого порядка — I(1). Если $|a| < 1$, то ряд стационарный — I(0).

   AR(1) можно записать в виде:

   $$\Delta y_t = b * y_{t-1} + \varepsilon_t \tag{4}$$

   где $b = a - 1$, а $\triangle$ — оператор разности первого порядка $\triangle y_t = y_t - y_{t-1}$. Поэтому проверка гипотезы о единичном корне в данном представлении означает проверку нулевой гипотезы о равенстве нулю коэффициента b.

3. Расширенный тест Дики — Фуллера: Если в тестовые регрессии добавить лаги первых разностей временного ряда, то получем расширенную модель Дики-Фуллера.
   Необходимость включения лагов первых разностей связана с тем, что процесс может быть авторегрессией не первого, а более высокого порядка.

$$y_t = a_1 * y_{t-1} + a_2 * y_{t-2} + \varepsilon_t \tag{5}$$

   Данную модель можно представить в виде:

$$\Delta y_t = (a_1 + a_2 - 1) * y_{t-1} - a_2 * \Delta y_{t-1} + \varepsilon_t \tag{6}$$

   Если временной ряд имеет один единичный корень, то первые разности по определению стационарны. А поскольку y_{t-1} по предположению нестационарен, то если коэффициент при нём не равен нулю, уравнение противоречиво. Таким образом, из предположения об интегрированности первого порядка для такого ряда следует, что $a_1 + a_2 - 1 = 0$.

4. Автокорреляционная функция(АКФ): Автокорреляционная функция — зависимость взаимосвязи между функцией (сигналом) и её сдвинутой копией от величины временного сдвига. Для детерминированных сигналов автокорреляционная функция (АКФ) сигнала f(t) определяется интегралом:

$$\int_{-\infty}^{\infty} f(t) * f(t - \tau) * dt \tag{7}$$

5. АРИМА Модель $ARIMA(p, d, q)$ для нестационарного временного ряда X_t имеет вид:

$$\triangle^d * X_t = c + \sum_{i=1}^{p} a_i * \triangle^d * X_{t-1} + \sum_{j=1}^{q} b_i * \varepsilon_{t-j} + \varepsilon_t \tag{8}$$

   , где ε_t - стационарный временной ряд; c, a_i, b_j - параметры модели. $\triangle^d$ - оператор разности временного ряда порядка d (последовательное взятие d раз разностей первого порядка - сначала от временного ряда, затем от полученных разностей первого порядка, затем от второго порядка и т.д.)

Описание программы

С помощью matplotlib строится график данного временного ряда. Далее осуществляется проверка на стационарность временного ряда. С помощью функций Series.rolling.mean() и Series.rolling.std() находятся среднее и стандартное отклонения соответственно, строятся их графики вместе с оригинальным, а затем проводится тест Дики-Фуллера с помощью функции adfuller из модуля pandas.

Разлагаем ряд на тренд, сезонность и остаток с помощью функции из модуля statmodels seasonal_decompose с параметрами model = 'additive' и model = 'multiplicate' для аддитивной и мультипликативной моделей соответственно. Далее идет проверка на стационарность рядов описанным выше способом.

Ряд проверяется на интегрируемость порядка k. Поэтому для проверки стационарности проведем обобщенный тест Дики-Фуллера на наличие единичных корней. Для этого в модуле statsmodels есть функция adfuller(). Если проведенный тест подтвердил предположения о не стационарности ряда, для нахождения k берется разность рядов. Если первые разности ряда стационарны, то он называется интегрированным рядом первого порядка. В нашем случае k = 1 и ряд интегрируем, значит с помощью функций автокорреляции и частичной автокорреляции подбираются параметры для модели ARIMA. Для построения модели нужно знать ее порядок, состоящий из 3-х параметров: p — порядок компоненты AR, d — порядок интегрированного ряда, q — порядок компонетны MA.

Параметр d равен 1, осталось определить p и q. Для их определения надо изучить автокорреляционную(ACF) и частично автокорреляционную(PACF) функции для ряда первых разностей. ACF поможет определить q, т. к. по ее коррелограмме можно определить количество автокорреляционных коэффициентов сильно отличных от 0 в модели MA. PACF поможет определить p, т. к. по ее коррелограмме можно определить максимальный номер коэффициента сильно отличного от 0 в модели AR. Чтобы построить соответствующие коррелограммы, в пакете statsmodels имеются функции: plot_acf() и plot_pacf(). Они выводят графики ACF и PACF, у которых по оси X откладываются номера лагов, а по оси Y значения соответствующих функций. В первой модели ACF экспоненциально затухает, начиная с первого лага, причем затухание может носить монотонный или колебательный характер. PACF затухает экспоненциально, монотонно или колебательно. Это означает, что p = 1, а q = 3. Во второй модели мы ссылаемся на стандартный подход по выбору параметров: q - номер последнего лага, при котором автокорреляция значима, p - номер последнего лага при котором частичная автокорреляция значима. Получаем p = 12, q = 3. p =1 , q = 4 выбираем по тому же принципу.

Далее строятся модели ARIMA и осуществляется прогноз. Строится график, на котором изображены данные из файла testing.xlsx и построенный прогноз для каждой из моделей. Для каждой модели считается R2 - коэффициент детерминации, чтобы понять какой процент наблюдений описывает данная модель, и критерий Акаике (AIC), выбирающий наилучшую модель.

Выводы

Под стационарностью понимают свойство процесса не менять своих статистических характеристик с течением времени, а именно постоянство математического ожидания, постоянство дисперсии и независимость ковариационной функции от времени (должна зависеть только от расстояния между наблюдениями). Оригинальный ряд и трендовая составляющая для обеих моделей не стационарны, т.к. стандартное отклонение и скользящее среднее зависят от времени. Тест Дики-Фуллера подтверждает данный вывод.

Сезонная составляющая и остаток (для обеих моделей) стационарны, их стандартное отклонение и скользящее среднее не зависят от времени.

Если r2 score близок к 1, то условная дисперсия модели достаточно мала и весьма вероятно, что модель неплохо описывает данные. Если же он сильно меньше 1, то с большей долей уверенности модель не отражает реальное положение вещей. Значения наших моделей равны: -3.32, 0.02, -3.18. Это говорит о том, что они не являются точными. Считается, что наилучшей будет модель с наименьшим значением критерия Акаике, и в нашем случае это модель ARIMA с значением 246.56. Однако погрешность этой модели велика, следовательно мы считаем модель со значением 248.79 наилучшей в силу меньшей погрешности вычислений.

Необходимые компоненты

- Библиотеки
  - matplotlib - пакет, используемаый для отрисовки графиков
  - statsmodels - пакет Python, который позволяет пользователям исследовать данные, оценивать статистические модели и выполнять статистические тесты. Он дополняет модуль статистики SciPy. Мы используем ее для проведения теста Дики-Фуллера, а так же для построения модели ARIMA.
  - sklearn - пакет для машшиного обучения. Мы используем ее для оценки $r2\_value$.
  - pandas - библиотека предназначенная для хранения таблиц. Так же содержит огромное количество универсальных функций для их комфортной обрабоки.
  - pylab - большой универсальный пакет питон. Мы используем для задания параметров отрисовки.
- Программы
  - Jupyter Notebook

Участники

Данько Артем - README.pdf
Ковалева Василина - задание 1, 3
Виль Вадим - задание 2